Laara E, Day NE, Hakama M. Trends in mortality from cervical cancer in the Nordic countries: association with organised screening programmes. *Lancet* 1987;i:1247-9.
McPherson A, Austoker J. Cervical cytology. In: McPherson A, ed. *Women's problems in general practice*. 3rd ed. Oxford: Oxford University Press, 1992: 227-52.
Meadows P. Study of women overdue for a smear test in a general practice cervical screening programme. *J R Coll Gen Pract* 1987;37:500-3.
Nathoo V. Investigation of non-responders at a cervical cancer screening clinic in Manchester. *BMJ* 1988;**296**:1041-2.
Nichols S. Women's preferences for sex of doctor: a postal survey. *J R Coll Gen Pract* 1987;**37**:540-3.
Pierce M, Lundy S, Palanisamy A, Winning S, King J. Prospective randomised controlled trial of methods of call and recall for cervical cytology screening. *BMJ* 1989;**299**:160-2.
Shafi M, Luesley D, Jordan J. Mild cervical cytological abnormalities. *BMJ* 1992;**305**:1040-1.
Shroff K, Corrigan AM, Bosher M, Edmonds MP, Sacks D, Coleman DV. Cervical screening in an inner city area: response to a call system in general practice. *BMJ* 1988;**297**:1317-8.
Soutter WP, Fletcher A. Invasive cancer of the cervix in women with mild dyskaryosis followed up cytologically. *BMJ* 1994;**308**:1421-3.
Wilkinson C, Jones J, McBride J. Anxiety caused by abnormal result of cervical smear test: a controlled trial. *BMJ* 1990;**300**:440.
Wilkinson C. Abnormal cervical smear test results: old dilemmas and new directions. *Br J Gen Pract* 1992;**42**:336-9.
Wilson A, Leeming A. Cervical cytology screening: a comparison of two call systems. *BMJ* 1987;**295**:181-2.

A complete list of references is available from the author.

Statistics Notes

One and two sided tests of significance

J Martin Bland, Douglas G Altman

This is the eighth in a series of occasional notes on medical statistics.

In some comparisons—for example, between two means or two proportions—there is a choice between two sided or one sided tests of significance (all comparisons of three or more groups are two sided).

When we use a test of significance to compare two groups we usually start with the null hypothesis that there is no difference between the populations from which the data come. If this hypothesis is not true the alternative hypothesis must be true—that there is a difference. Since the null hypothesis specifies no direction for the difference nor does the alternative hypothesis, and so we have a two sided test. In a one sided test the alternative hypothesis does specify a direction—for example, that an active treatment is better than a placebo. This is sometimes justified by saying that we are not interested in the possibility that the active treatment is worse than no treatment. This possibility is still part of the test; it is part of the null hypothesis, which now states that the difference in the population is zero or in favour of the placebo.

Paired t test analyses of difference in forced vital capacity (ml) between first and subsequent visits (n=25)

	Change in forced vital capacity from baseline		
	1 week	3 months	1 year
Mean	48	−63	−49
Standard error	42	33	55
P (one-sided)	0·13	0·032	0·19

A one sided test is sometimes appropriate. Luthra *et al* investigated the effects of laparoscopy and hydrotubation on the fertility of women presenting at an infertility clinic.[1] After some months laparoscopy was carried out on those who had still not conceived. These women were then observed for several further months and some of these women also conceived. The conception rate in the period before laparoscopy was compared with that afterwards. The less fertile a woman is the longer it is likely to take her to conceive. Hence, the women who had the laparoscopy should have a lower conception rate (by an unknown amount) than the larger group who entered the study, because the more fertile women had conceived before their turn for laparoscopy came. To see whether laparoscopy increased fertility, Luthra *et al* tested the null hypothesis that the conception rate after laparoscopy was less than or equal to that before. The alternative hypothesis was that the conception rate after laparoscopy was higher than that before. A two sided test was inappropriate because if the laparoscopy had no effect on fertility the conception rate after laparoscopy was expected to be lower.

One sided tests are not often used, and sometimes they are not justified. Consider the following example. Twenty five patients with breast cancer were given radiotherapy treatment of 50 Gy in fractions of 2 Gy over 5 weeks.[2] Lung function was measured initially, at one week, at three months, and at one year. The aim of the study was to see whether lung function was lowered following radiotherapy. Some of the results are shown in the table, the forced vital capacity being compared between the initial and each subsequent visit using one sided tests. The direction of the one sided tests was not specified, but it may appear reasonable to test the alternative hypothesis that forced vital capacity decreases after radiotherapy, as there is no reason to suppose that damage to the lungs would increase it. The null hypothesis is that forced vital capacity does not change or increases. If the forced vital capacity increases, this is consistent with the null hypothesis, and the more it increases the more consistent the data are with the null hypothesis. Because the differences are not all in the same direction, at least one P value should be greater than 0·5. What has been done here is to test the null hypothesis that forced vital capacity does not change or decreases from visit 1 to visit 2 (nine week), and to test the null hypothesis that it does not change or increases from visit 1 to visit 3 (three months) or visit 4 (one year). These authors seem to have carried out one sided tests in both directions for each visit and then taken the smaller probability. If there is no difference in the population the probability of getting a significant difference by this approach is 10%, not 5% as it should be. The chance of a spurious significant difference is doubled. Two sided tests should be used, which would give probabilities of 0·26, 0·064, and 0·38, and no significant differences.

In general a one sided test is appropriate when a large difference in one direction would lead to the same action as no difference at all. Expectation of a difference in a particular direction is not adequate justification. In medicine, things do not always work out as expected, and researchers may be surprised by their results. For example, Galløe *et al* found that oral magnesium significantly increased the risk of cardiac events, rather than decreasing it as they had hoped.[3] If a new treatment kills a lot of patients we should not simply abandon it; we should ask why this happened.

Two sided tests should be used unless there is a very good reason for doing otherwise. If one sided tests are to be used the direction of the test must be specified in advance. One sided tests should never be used simply as a device to make a conventionally non-significant difference significant.

Department of Public Health Sciences, St George's Hospital Medical School, London SW17 0RE
J Martin Bland, *reader in medical statistics*

Medical Statistics Laboratory, Imperial Cancer Research Fund, London WC2A 3PX
Douglas G Altman, *head*

BMJ 1994;309:248

1 Lund MB, Myhre KI, Melsom H, Johansen B. The effect on pulmonary function of tangential field technique in radiotherapy for carcinoma of the breast. *Br J Radiol* 1991;**64**:520-3.
2 Luthra P, Bland JM, Stanton SL. Incidence of pregnancy after laparoscopy and hydrotubation. *BMJ* 1982;**284**:1013.
3 Galløe AM, Rasmussen HS, Jørgensen LN, Aurup P, Balsløv S, Cintin C, Graudal N, McNair P. Influence of oral magnesium supplementation on cardiac events among survivors of an acute myocardial infarction. *BMJ* 1993;**307**:585-7.